DHS Data Editing and Imputation

Trevor Croft, Demographic and Health Surveys

Abstract

Producing good quality demographic and health data and making it accessible to data users worldwide is one of the main aims of the Demographic and Health Surveys program. However, large scale surveys in developing countries, particularly those collecting retrospective data, are prone to poor reporting. Survey data have suffered traditionally from incomplete and inconsistent reporting, To handle these problems, DHS performed extensive data editing operations. In addition, imputation procedures were established to deal with partial reporting of dates of key events in the respondent's life.

General techniques for handling incomplete and inconsistent data are considered. The paper then presents the DHS approach to data editing. The major focus is on the editing of dates of events and the intervals between events. The editing and imputation process starts with the calculation of initial logical ranges for each date, and gradually constrains these ranges to produce final logical ranges. Inconsistent data are reported in error listings during this process. Dates are imputed for events with incomplete reporting within these final logical ranges. The levels of imputation required in the DHS-I surveys are presented.

Various problem areas involved with the imputation of incomplete dates are explained. These include biases caused by questionnaire design, miscalculation of dates by interviewers and ancillary data biases. Problems relating to fine temporal variables and to unconstrained ranges for dates are also reviewed.

Finally, the changes introduced as part of the editing and imputation procedures for DHS-II to resolve some of the problem areas are presented and ideas for further improvements are discussed.

DHS Data Editing and Imputation Trevor Croft,

Demographic and Health Surveys

I. Introduction

A. Objectives of the paper

One of the primary goals of the Demographic and Health Surveys (DHS) program is to produce high-quality data and make it available for analysis in a coherent and consistent form. Demographic surveys in developing countries are prone to incomplete or partial reporting of responses. Additionally, complex questionnaires inevitably allow scope for inconsistent responses to be recorded for different questions (Otto 1987). For the analyst this results in a data file containing incomplete or inconsistent data, complicating the analysis considerably. In order to avoid these problems, the DHS program has adopted a policy of editing and imputation which results in a data file that accurately reflects the population studied and may be readily used for analysis.

The objective of this paper is to present the DHS approach to the production of a final edited data file, focusing on the editing of dates of events and the imputation of incomplete dates. The paper will discuss various approaches to the problems of partial and inconsistent data, and the need for procedures to handle this data, and will then present the DHS approach to editing and imputation. The results of this approach will be discussed and the problems found in the use of these procedures will be presented. Finally, the paper will discuss the changes made in the procedures for editing and imputation for phase II of the Demographic and Health Surveys program and consider further improvements that may be made to these procedures.

B. Background

1. Data Editing approaches

Two of the major sources of problems in DHS surveys and in many other survey programs are partial or incomplete reporting of information, and inconsistent responses to different questions in the survey. Consideration must first be made of how to handle these problems within the data file. For missing data there are three obvious possibilities:

- Leave the question blank.

  In general, this is not an ideal solution as responses to questions that are skipped due to the flow of the questionnaire are usually left blank. However, if a special code is given to skipped questions, then a blank field for a missing question could be considered as a special code for missing data, and thus a particular case of the next option.

- Assign a special code for the question indicating that there was no response in the questionnaire.

  This approach allows the fact that no response was reported or recorded to be registered in the data file and permits cases with missing data to be handled specially either during data analysis or during an imputation stage. It also permits decisions as to how these data are to be handled to be postponed until all data, or at least a significant amount of data, are collected, when decisions may be based on the data encountered and not on abstract ideas.

• Deduce a response for the question.

This is not recommended as it could lead to biases in the data. In most situations there would be no rationale as to how a particular response should be chosen. However, in some situations it is possible, and indeed desirable, to deduce the response from other information in the questionnaire. This is particularly true for questions which affect the flow of the questionnaire. For example, if the question 'Are you pregnant now?' was left blank, but the following question 'How many months pregnant are you?' contained the response '3 months', the response to the First question can be deduced to be 'Yes'. This response should be used in such a case as using a special code for the missing data would generally be assumed to be a negative response and the following question may be skipped in data entry.

Some inconsistent responses may be found during data entry, but it is more usual for a secondary editing phase to uncover the inconsistent data. In dealing with these data there are similar possibilities:

• Leave the data. unchanged.

For some questions two pieces of data may appear to be inconsistent but if there is no item which is obviously incorrect or the inconsistency has no practical effect on the analysis then the data may be left unchanged.

• Give a special code to indicate the response was inconsistent with other information reported.

This has the advantage that the analyst will not have to deal with inconsistent data during analysis, thus simplifying the analysis. The disadvantage is that the original data has been lost (although it was assumed to have been incorrect). A modified approach to this is to use a separate flag variable to indicate that the question was found to be inconsistent, coding the reason for the inconsistency in the flag variable, but leaving the original variable unchanged. The main disadvantages to this are that a plethora of flag variables may be added to the data file, and the analyst will have to take the flag variable into account as well as the data variable in any analysis, thus further complicating analysis.

• Deduce a response for the question.

For certain questions it is sometimes possible to deduce the correct response from other responses in the questionnaire. In general, responses would only be changed to another response when there is convincing evidence that the new response is correct.

A mix of the three approaches in editing data for consistency is often used, and different variables merit different treatment.

2. Rationale for imputation

There are various approaches to the editing of data for completeness and consistency that range from doing nothing at all to completely whitewashing the data to remove all inconsistencies and produce complete data for every case and variable. In general, the optimal solution lies somewhere in between, for obvious reasons:

- Unedited data requires careful handling and laborious checking of data during analysis to avoid misleading conclusions. Incomplete data requires complex control of all possible combinations of complete and incomplete responses.

- Whitewashed data in which all inconsistencies have been removed and all incomplete data have been imputed may produce biased results. The cost in terms of time and effort to produce a 'clean' data file of this kind is also prohibitive.

Certain variables are central to the analysis of the data and for this reason incomplete or inconsistent data may not be tolerated for these questions. It should be noted that some responses which are deemed acceptable during interviewing, are not desirable when analyzing data. For example, in response to the question 'In what year and month were you born?' the year of birth may be known, but the month of birth is unknown. From the analysis viewpoint, this information is considered to be partial data. In fact, any variable assigned a special code for inconsistent data may also be considered to contain partial data.

In many cases there are related variables that provide sufficient information to deduce the correct response to key variables. However, for some of these variables, it may not be possible to deduce the correct response to a question where missing or partial data has been given or where the original response was inconsistent with other information, but it is necessary that the question contains a valid response. In these situations the responses to variables may be imputed.

3. Methods of imputation

Imputation is the process of attributing a characteristic to a case based on known characteristics of a population in general and other particular characteristics of the case in question. In other words, a response to a question is imputed, based on a set of rules which may take into account responses to other questions for the same case, and responses to questions in other cases in the population.

There are really two steps in imputation: Firstly, a set of rules are used to restrict the possible responses that may be attributed to the case; and secondly, a method of choosing a response is applied when no further reduction of the set of acceptable responses is possible.

For the first step, the rules can range from no restrictions at all in the simplest case, to extremely elaborate constraints. For the second step there are four major methods:

- Cold deck

  Imputation within prescribed constraints based on a predefined distribution of cases from the population, usually from a separate source, taking into account certain characteristics of the case in question.

- Hot deck

  Imputation within prescribed constraints based on responses from earlier cases processed in the population who have certain characteristics in common with the case in question.

- Random

  Imputation within the prescribed constraints is performed randomly. Usually the randomly imputed cases will be distributed uniformly within the constraints, however some algorithms may call for normally distributed random responses.

- Midpoint

  Imputation within the prescribed constraints is performed, by selecting the midpoint of the range of acceptable responses. This form of imputation would only be used with continuous variables and not with categorical variables.

In census data processing, many of the variables may be imputed to avoid incomplete reporting or inconsistent information. This simplifies the analysis stage of the census and produces little bias in the results, as the number of cases to be imputed for each variable is generally a small proportion of the total number of cases. Most census applications tend to use the hot deck approach, imputing values based on the values found in earlier cases with a few constraints on the set of acceptable responses based on the responses to other questions from the same case.

In contrast, the World Fertility Survey (WFS) used imputation for a select set of key variables (Trussell 1987). The constraints put on the set of acceptable responses for the variables were extremely elaborate and generally restricted these responses to a narrow range. Within this constrained range a response was usually selected randomly, although some surveys used midpoint imputation. The key variables that were imputed were:

- Date of birth and age of the respondent
- Date of birth and age of a child
- Date of beginning or ending a union
- Date of sterilization
- Date of expected delivery of current pregnancy

Partial or missing data on most variables causes a minor inconvenience for the analyst, but does not prevent analyses from being performed; however, missing or partial data for these key variables would have severely affected analyses involving these variables. When an analyst is confronted with missing data he has the choice of discarding the case from the analysis or effectively performing his own imputation of the data. As will be shown in the next section, discarding cases with incomplete data would seriously bias any analysis. On the other hand, the inter-relationships of each of the key variables and many other related variables make the imputation of consistent data a complicated task. In addition, if each analyst was performing his own imputation of these data, it would be unlikely that any other analyst would be able to reproduce the same results. For these reasons WFS adopted a policy of imputing these key variables as part of the editing process, following standard procedures (WFS 1980, Otto 1980).

4. WFS level of complete dates

Table I indicates the level of complete reporting of dates of events in the WFS surveys. These range from a handful of cases in Yemen with both month and year of the respondent's birth recorded to complete reporting of the dates of all events in the Korean survey. Levels of reporting were high in the Latin American/Caribbean and some Asian surveys, but consistently lower in the African, Near Eastern and Indian sub-continent surveys. Events occurring closer to the interview

Table I. Percentage of events in which both month and year or event were reported in WFS surveys

Country	Respondent's birth	First union	Birth of all children	Birth of first child	Birth of Last child
Sub-Saharan Africa					
Benin	9.2	4.9	12.4	14.9	26.8
Cameroon	28.3	21.0	40.9	42.0	56.8
Cote D'Ivoire	20.3	12.2	28.4	29.4	56.6
Ghana	52.1	40.3	63.5	64.2	78.3
Kenya	33.6	68.9	75.4	78.1	86.5
Lesotho	72.5	88.2	89.7	91.6	94.3
Mauritania	3.9	7.4	11.6	12.5	19.8
Nigeria	15.8	19.1	26.8	27.8	36.9
Senegal	38.2	69.4	99.0	98.8	99.3
Sudan	21.5	41.1	63.0	60.3	83.8
North Africa/Near East					
Egypt	26.2	36.8	41.4	45.4	57.5
Jordan	29.7	58.4	66.5	69.0	84.2
Morocco	22.2	35.2	59.7	58.6	69.2
Syria	57.3	79.0	83.2	83.0	95.2
Tunisia	88.2	53.3	70.4	71.1	75.2
Yemen AR	0.3	7.6	11.0	10.4	40.3
Asia					
Bangladesh	1.4	11.4	12.3	14.8	32.6
Fiji	67.6	85.3	86.3	88.3	96.1
Indonesia	22.3	45.7	46.5	50.8	55.5
Korea, Rep.	100.0	100.0	100.0	100.0	100.0
Malaysia	57.0	61.8	86.2	82.2	95.1
Nepal	13.4	27.3	99.9[a]	100.0[a]	100.0[a]
Pakistan	6.8	73.2	79.8	79.1	90.1
Philippines	97.3	95.8	96.2	98.4	99.0
Sri Lanka	67.0	70.3	73.4	77.8	83.3
Thailand	85.0	75.3	84.2	87.3	90.7
Latin America/Caribbean					
Colombia	100.0[b]	100.0[b]	100.0[b]	100.0[b]	100.0[b]
Costa Rica	100.0[b]	100.0[b]	100.0[b]	100.0[b]	100.0[b]
Dominican Rep.	85.9[c]	73.3[c]	91.0[c]	100.0[b]	100.0[b]
Ecuador	99.9	66.8	78.5	85.0	89.3
Guyana	98.0	78.7	91.2	95.0	93.4
Haiti	91.7	92.7	93.8	94.5	96.5
Jamaica	94.6	53.0	90.6	92.6	93.2
Mexico	100.0[b]	100.0[b]	100.0[b]	100.0[b]	100.0
Panama	99.4	94.5	97.7	98.2	98.9
Paraguay	99.9	98.0	99.7	99.7	99.9
Peru	94.7	81.2	93.1	95.0	97.9
Trinidad & Tob.	98.3	100.0[b]	94.5	96.2	96.3
Venezuela	100.0[b]	100.0[b]	100.0[b]	100.0[b]	100.0[b]

[a] The birth history automatically imputes only calendar year for all births
[b] After imputation – raw data file not available
[c] Figures are those reported in the First Country Report

Source: Chidambaram and Sathar, 1984.

date were predictably better reported in almost all surveys.

As shown in Chidambaram and Sathar, 1984, women with a higher level of education and women who live in major urban areas were more likely to report the dates of events completely, Clearly the option of dropping cases with incomplete dates of events from the analysis is not feasible due to the biases that would be introduced. Imputation is the only realistic solution.

II. DHS-I Approach to Editing of Data

The DHS approach to editing of data foresees three distinct phases at which data are machine edited. Firstly during data entry, secondly as a secondary editing stage and thirdly the imputation stage. Each stage is a distinct entity with data passing to the secondary editing stage only after all questionnaires from a sample point have been entered. Similarly, the imputation stage is reached only after all questionnaires have been entered and edited. It should be noted that questionnaires are also manually edited, both during the fieldwork and in the central office prior to data entry, but that editing is usually limited to a more rudimentary check for certain basic problems in the questionnaire.

A. Editing during Data Entry

During data entry, editing of data is restricted to controlling the structure of the data file, the skip patterns through the questionnaire, the range of valid values for each variable and the consistency of certain variables as they relate to maintaining the structural integrity of the data. DHS uses a special code consisting of a field of 9s to indicate data that was missing on the questionnaire. For example, the missing value for a two digit field is 99. For most consistency edits at this stage there are sufficient data to indicate the correct response when items of information are inconsistent and responses can be deduced. A second type of check is often included at data entry stage to attempt to identify keying errors. No correction of data is required in these cases if the data on the questionnaire has been correctly entered.

B. Secondary Data Editing

During the secondary editing stage complex checks are introduced to verify the internal consistency of information throughout the questionnaire. The editing and correction rules are described in detail in the DHS-I Data Processing Manual. In general, the corrections are of two types:

- Assignment of a special code indicating inconsistent data. DHS uses a field of 9s with a 7 as the last digit. For example, the inconsistent code for a three digit field is the value 997.
- Deduction of a 'correct' response from related questions and other information found in the questionnaire.

The editing rules allow responses to be changed to another valid response only in situations where it is clear that the new response is correct. In other situations, the inconsistent item of data is changed to the 'inconsistent' code. One special rule, known as the 'rule of one', is used for some questions. This rule allows a response to be changed by one unit (one month or one year depending on the unit of response to the question) if the modified response will be consistent with the other related data.

C. Imputation

The third stage of the editing process is the date imputation stage, in which a new data file is produced, containing complete dates of each event reported in the questionnaire. The events for which DHS imputes dates are as follows:

- Date of birth of the respondent
- Date of first union
- Date of birth of each child
- Date of conception of current pregnancy
- Date of sterilization of respondent or partner

In this stage, partial or incomplete dates are imputed from the known related information. Dates that are partial include dates for which no month or no year was reported, either because the questionnaire was blank (missing data), the information given was not consistent with other information (inconsistent data), or because the respondent did not know the exact date (unknown data).

III. DHS-I Date Editing and Imputation

A. General overview

The approach to editing dates of events and the approach to the imputation of missing dates consistent with other reported data is identical and, in fact, the same program is used for both the secondary editing stage and the imputation stage. The only difference is that a new data file is created in the imputation stage and this data file contains imputed dates for each event reported.

The machine imputation of incomplete dates is the final step in the process of editing dates for consistency. The method used in the DHS program for imputation of dates relies on the construction of logical ranges for each date, which are refined in three steps resulting in successively narrower or constrained ranges. At each step in the process, inconsistencies are found and error messages displayed to allow for the data to be corrected. If no errors are found in the data then, in a fourth step, values are randomly assigned from within the final constrained logical ranges, to be used as the imputed dates.

At the first stage, an unconstrained range is constructed from the available information. If month and year are both available, the upper and lower bounds of the range will be identical. If only a year is available, the unconstrained range will span 12 months. If no year is given the unconstrained range will cover the full range of possible dates, i.e. 50 years before interview until 15 years before interview for the date of birth of the respondent.

At the second stage, ranges are adjusted in the light of isolated constraints. These are constraints affecting only the date in question. An example is the constraint induced upon date of birth of the respondent by information on her current age.

At the third stage, ranges are adjusted to satisfy neighboring constraints. These are constraints arising from the fact that some dates form a logical sequence with required minimum intervals between them. An example is the constraint induced on the date of birth of a child by the date of the previous birth and the minimum required time from conception to delivery.

At the end of this process the difference between the upper bound and the lower bound of the range may be either negative, zero or positive:

1) If the range is negative, the date is inconsistent, in the sense that it violates an isolated or neighboring constraint, and it should be corrected manually, as part of the secondary editing process. Note that the error may be in the date in question, a neighboring date or in some ancillary data related to the event.

2) If the range is zero, the date is consistent with other related information and known, either because it was fully specified in the first place or because it was fully constrained, i.e. the constraints were sufficient to restrict it to a single month.

3) If the range is positive, the date is consistent with other related information but incomplete, because the constraints were not sufficient to restrict it to a single month. In this case the date will be imputed choosing a random date uniformly distributed within the logical range.

B. Logical ranges

The first step in the imputation process is to construct logical ranges for the dates of each event. These ranges are the earliest and latest possible dates at which the event can have occurred, based on the reported month and year of the event. No other information is taken into account at this stage. If both the month and year of the event are reported the lower and upper bounds of the range are the same.

The logical ranges for each of the events recorded in the questionnaire are built up into an event table. The event table facilitates the checking of consistency of date and interval responses in the questionnaire. All dates in the event table are recorded in century month codes (CMCs). The century month code for a date is the number of months since the start of the century. For example, January 1900 is month 1, December 1900 is month 12, January 1901 is month 13 and August 1991 is month 1100. To calculate the century month code for the date of an event, simply multiply the year of the event by 12 and add the month of the event. The century month code is used throughout the date editing and imputation process as it considerably simplifies the calculations involved.

Table II. Event Table and Initial Logical Ranges

	Event	Date	Initial Ranges
1.	Birth of the respondent	DK-DK	501-920
1	First Union	DK-81	973-984
3.	Child 1	DK-DK	621-1100
4.	Child 2	02-Missing	621-1100
5.	Child 3	08-86	1040-1040
6.	Child 4 (twin)	DK-87	1045-1056
7.	Child 5 (twin)	DK-87	1045-1056
8.	Child 6	08-89	1076-1076
9.	Current Pregnancy		1090-1099
10.	Interview	08-91	1100-1100

The event table is also used to record other information related to the event, such as the type of the event, and data relating to intervals between events. Table II gives an example for one respondent of the data used in initially creating the event table. The table shows typical data recorded for the dates of each event. For the current pregnancy no date of conception is asked in the questionnaire, just the duration of the pregnancy, which is used as an isolated constraint. The initial logical ranges here are created solely based upon the dates given. Other information, such as age of the respondent and her children, are used in the steps that follow. In practice the date of first union is not held as part of the event table, in order to allow for pre-marital births, but is handled separately.

C. Isolated constraints

The second step in the date editing and imputation involves narrowing the bounds for the dates of each event using isolated constraints. Isolated constraints are items of data relating to a particular event, but with no relation to any other event (other than the date of interview, which is always fully specified). Most isolated constraints apply a constraint upon both the lower bound and the upper bound of a date. The age of the respondent and age of each child are isolated constraints upon the date of birth of the respondent and the date of birth of each child respectively. Duration of current pregnancy is an isolated constraint on the date of conception of the pregnancy. Age at death of children, and dates of vaccination are also isolated constraints on the dates of birth of children, but they only constrain the upper bound of the dates.

Table III. Data Used for Initial Ranges and Isolated Constraints

Events	Initial Range	Isolated Constraints
Birth of respondent	Year and month	Age
First union	Year and month	Age at union
Birth of child	Year and month	LIVING: Age of child Dates of vaccination Existence of health data DEAD: Age at death
Conception of Current Pregnancy	-	Duration of pregnancy
Sterilization	Year and month	-
Interview	Year and month	-

The age of the respondent at first union is also treated as an isolated constraint as it constrains both the lower and upper bound of the date of first union, but as this constraint is relative to the date of birth of the respondent it is clearly not truly an <u>isolated</u> constraint. Age at first union was only asked if the year of first union was not known. Table III gives a summary of the isolated constraints used in DHS-I.

The constraints induced by these variables, must overlap by at least one month with the initial ranges for the data to be consistent. If there is no overlap, there is an inconsistency between the date of the event and the isolated constraint. This inconsistency would have to be resolved before the imputation step.

Table IV shows how the isolated constraints affect the range of possible dates for each event. The constrained ranges are produced by taking the maximum of the lower bound of the constraint and the lower bound of the initial range and the minimum of the upper bound of the constraint and the upper bound of the initial range. It should be noted that the logical range for any date is never increased in the process of producing the final logical ranges for imputation.

Table IV. Isolated Constraints

	Event	Date	Initial Ranges	Age	Constraints	Constrained Ranges
1.	Birth	DK-DK	501-920	27	765-776	765-776
2.	First Union	DK-81	973-984	Not asked		973-984
3.	Child 1	DK-DK	621-1100	Died 1 m	-1099	621-1099
4.	Child 2	02-Missing	621-1100	06	1017-1028	1017-1028
5.	Child 3	08-86	1040-1040	Died 3 m	-1097	1040-1040
6.	Child 4 (twin)	DK-87	1045-1056	Died 0 m	-1100	1045-1056
7.	Child 5 (twin)	DK-87	1045-1056	04	1041-1052	1045-1052
8.	Child 6	08-89	1076-1076	01	1076-1088	1076-1076
9.	Pregnancy		1090-1099	3 months	1097-1097	1097-1097
10.	Interview	08-91	1100-1100			1100-1100

Note that all ages are treated as the age in completed years. Some societies tend to round their ages up, but the questionnaire explicitly asks for age in terms of completed years ('How old were you at your last birthday?'). 'A thirteenth month is initially allowed on the lower end of the age constraints as the child may be expecting a birthday in the month of interview. This extra month is discarded if other months within the constraints are consistent with the initial range, in order to avoid the imputation of an age that is different from the age originally reported. However, for child 6, that month is the only month consistent with the initial date of birth reported. This is possible if the child was born on August 15th 1989 and the interview was on August 6th 1991. At the analysis stage this child will be treated as being two years old, based on its date of birth.

D. Neighboring constraints

Neighboring constraints are restrictions placed upon the range of acceptable dates by earlier and later events in the respondent's life. Neighboring constraints fall into two categories: minimum interval constraints and ancillary data constraints. Minimum intervals are defined between each event to ensure that events are no closer together than physically possible. Ancillary data provides additional information about the intervals between events and are used to enlarge the minimum possible interval between events.

The most obvious example of a minimum interval constraint is the gestation length of a pregnancy, which is usually nine months, implying that two births cannot be less than nine months apart. The earliest acceptable date for the birth of one child, plus the minimum interval between births, becomes a constraint on the lower bound of the date of birth of the following child. Similarly, the latest possible date of birth of a child, less the minimum birth interval, is a constraint on the upper bound of the date of birth of the preceding child. In practice, a minimum interval of seven months is used to allow for premature births.

Ancillary data which is used in addition to the minimum intervals, includes the durations of amenorrhea, abstinence and breastfeeding after the birth of a child, and the duration of contraceptive use in the interval prior to the birth of a child. The time since last sexual intercourse and the time since last menstrual period also provide constraints on the date of birth of the last child. Table V gives a summary of the minimum interval and ancillary data constraints used in DHS-I. The table is presented in terms of the constraints between two events.

The neighboring constraints are applied in a series of repetitions through the event table. Each type of constraint is applied to the intervals between the events, first in a forward direction and then in a backward direction. The process starts in the forward direction with the minimum interval between the birth of the respondent and the birth of the first child constraining the date of birth of the first child, the minimum intervals between births constraining the following births, the minimum interval between the last birth and conception of current pregnancy constraining the date of conception, and so on. The process is then reversed, starting with the last event, the date of interview, and applying the minimum interval constraints to the preceding event.

After applying the minimum interval constraints, the constraints for the durations of amenorrhea, abstinence, breastfeeding and contraceptive use are applied in turn, both in a forward direction and a backward direction. Finally the age at first sexual intercourse, time since last sexual intercourse and time since last menstrual period constraints are applied. The constraining of the interval between the date of first union and other events is usually handled separately.

If the upper bound constraint for a date produced by any part of this process is less than the lower bound for the date of the event, there is an inconsistency in the data. An error message is printed and correction would be made to the data. All such inconsistencies must be removed from the data before the final imputation stage.

Table VI gives an example of the effect of the neighboring constraints on the logical ranges produced by the isolated constraints stage. At each stage the logical ranges are updated by the newly constrained ranges to produce a gradually narrower range for the date of each event. When the minimum and maximum bound of the constraints overlap, i.e. the minimum is greater than the maximum, an inconsistency has been found in the data. In the example, the duration of amenorrhea when added to the minimum pregnancy duration exceeds the interval between the birth of child 3 and the birth of child 4. In this situation, the data would be corrected, probably by setting the duration of amenorrhea to the inconsistent code.

After applying the neighboring constraints, the ranges of dates for the birth of twins will be rationalized to produce identical logical ranges for each twin. Provided that there are no overlaps between the minimum and maximum possible dates of each event after all constraints are satisfied, the final two stages of the process will produce imputed dates for each event.

Table V. Data Used of Neighboring Constraints

Prior Event	Later Event	Minimum Interval	Ancillary Data
Birth of respondent	First union	10 years	Age at first sexual intercourse (if no information for date of first union)
	Birth of first child	12 years	Age at first sexual intercourse plus 7 months
	Conception of current pregnancy	12 years	Age at first sexual intercourse
	Sterilization	20 years	
	Interview	15 years	
First Union	Birth of first child	Non-negative (if no information for date of first union)	
	Sterilization	Non-negative	
	Interview	Non-negative	
Birth of child	Birth of child	7 months (0 months between twins of the same birth)	Duration of amenorrhea plus 7 months Duration of abstinence plus 7 months Duration of contraceptive use plus 7 months
	Conception of Current pregnancy	Non-negative	Duration of amenorrhea Duration of abstinence Duration of contraceptive use
	Sterilization	Non-negative	Duration of contraceptive use
	Interview	Non-negative	Duration of abstinence Duration of amenorrhea Duration of breastfeeding Duration of contraceptive use Time since last sex Time since last period
Conception of Current pregnancy	Interview	2 months	
Sterilization	Interview	Non-negative	

Table VI. Neighboring Constraints

Event	Logical Ranges	Min. Interval	Const-raints	Ameno-rrhea	Absti-nence	Breast-feeding	Ancillary Constraints	Constrained Ranges
1. Birth	765-776	-	-864				-864	765-776
2. First Union	973-984	120	885-1099				885-1099	973-984
3. Child 1	621-1099	0	973-1021				973-1021	973-1021
4. Child 2	1017-1028	7	980-1033				980-1033	1017-1028
5. Child 3	1040-1040	7	1024-1040	06	03	06	1024-1042	1040-1040
6. Child 4 (twin)	1045-1056	7	1047-1052	12	18	Never	[a]1053-1052	1050-1052
7. Child 5 (twin)	1045-1052	0	1047-1069	12	18	24	1047-1053	1047-1052
8. Child 6	1076-1076	7	1052-1097	08	12	18	1072-1078	1076-1076
9. Pregnancy	1097-1097	0	1083-1098				1095-1098	1097-1097
5. Interview	1100-1100	2	1099-				1099-	1100-1100

Age at first sexual intercourse	16 years
Time since last sexual intercourse	02 months
Time since last menstrual period	04 months

[a]The duration of amenorrhea plus the minimum interval of 7 months leads to an inconsistency between the lower and upper bound of the constraints. The final constrained range is based on the duration of abstinence plus the minimum interval of 7 months.

E. Gap increasing

After all constraints have been applied, the logical ranges of two events can overlap. If the dates of the events were imputed randomly within these ranges, it would be possible that the imputed date of an event might be after the imputed date of the following event. Since the goal of the imputation process is to produce dates for events which are consistent with each other and other ancillary data, it is necessary to ensure that no overlaps occur between the ranges for the dates of events and, indeed, that the required minimum intervals between events are preserved.

To increase the spacing between the events, half of the total of the amount of overlap between two events plus the minimum required interval between the events is subtracted from the upper bound of the earlier event and added to the lower bound of the later event. This is successively done for each of the intervals, to produce final logical ranges, which do not overlap and maintain the required minimum intervals between events.

F. Imputation within final range

DHS-I used a random imputation method to assign the imputed date within the final logical range for each event. The algorithm produced a uniform distribution of random numbers within the logical range. The resulting imputed dates were written to an output data file as part of an updated master data file. Dates were recorded in this data file in terms of century month codes. All analysis using dates of events make use of the century month code variables in the imputed data file rather than the original date variables. In addition, date flag variables for each event are written to this data file to indicate the original form in which the date was reported.

Table VII gives an example of the last steps in the imputation process, starting from the constrained ranges to produce the final logical ranges, with no overlapping ranges and minimum

Table VII. Final Logical Ranges and Imputed Dates

	Event	Constrained Ranges	Final Logical Ranges	Imputed CMCs	Imputed Dates
1.	Birth	765-776	765-776	770	02-64
2.	First Union	973-984	973-978	977	05-81
3.	Child 1	973-1021	979-1015	994	10-82
4.	Child 2	1017-1028	1023-1028	1025	05-85
5.	Child 3	1040-1040	1040-1040	1040	08-86
6.	Child 4 (twin)	1050-1052	1050-1052	1051	07-87
7.	Child 5 (twin)	1050-1052	1050-1052	1051	07-87
8.	Child 6	1076-1076	1076-1076	1076	08-89
9.	Pregnancy	1097-1097	1097-1097	1097	05-91
10.	Interview	1100-1100	1100-1100	1100	08-91

intervals between events preserved. The imputed dates are randomly assigned within the final logical ranges.

IV. Level of imputation in DHS-I surveys

Table VIII presents the level of imputation in DHS-I surveys for each of the events for which dates were imputed. In general, levels of complete reporting are highest in Latin America and the Caribbean, where over 95 percent of all birth dates are fully specified. In contrast most Sub-Saharan African countries have much lower levels of complete reporting. Countries in Africa and Asia that are in the midst of fertility transition are varied in the degree of complete reporting of dates.

The survey in Morocco requires special consideration as seasonal reporting was also allowed for each date. Each season was taken to represent three months of the year, although there is clearly some overlap between seasons. The figures presented in table VIII do not include the seasonal reporting as complete reporting, but reporting of year and season accounted for 3 percent of the respondent's birth dates, 60 percent of the dates of first union, 40 percent of the children's birth dates and 18 percent of the sterilization dates in Morocco.

A. Date of birth of respondent

As part of the data collection procedures used in DHS surveys, the interviewers are required to record the age of each respondent. As eligible women are selected for the individual interview on the basis of their age recorded in the household schedule, reporting of age of the respondent was almost universal. Occasionally it was necessary for the interviewer to estimate the age of the respondent during the interview, but in practice this is fairly rare.

Because the reporting of the age of the respondent was practically universal (there are only a handful of respondents throughout all of the DHS-I surveys with no age recorded), the imputation of date of birth of the respondent is restricted to a 12 month range in the worst case and to a narrower range if the year of birth of the respondent was also recorded.

Table VIII. Percentage of events in which both month and year of event were reported in DHS-I surveys							
Country	Respondent's birth	First Union	Birth of all children	Birth of first child	Birth of last child	Conception of pregnancy	Steril-ization
Sub-Saharan Africa							
Botswana	84.8	68.4	96.9	97.6	99.1	99.0	98.1
Burundi	38.2	59.8	79.7	80.6	93.2	99.3	-
Ghana	48.7	29.3	753	77.6	88.1	99.6	-
Kenya	63.4	81.1	96.5	97.2	98.2	99.1	-
Liberia	42.3	32.1	85.2[a]	85.5[a]	91.0[a]	99.5	-
Mali	9.0	6.3	34.9	33.5	53.8	98.8	-
Ondo State	65.8	62.7	99.9[a]	99.9[a]	100.0[a]	98.8	-
Senegal	34.1	16.6	76.5	73.7	92.2	99.4	-
Sudan	16.0	36.1	53.8	58.7	73.0	99.9	-
Togo	26.9	19.6	50.0	51.3	73.8	99.7	-
Uganda	74.9	86.8	99.9	99.9	99.9	99.7	-
Zimbabwe	89.9	76.9	99.4	99.4	99.8	99.7	97.4
North Africa/Near East							
Egypt	43.0	50.5	63.7	69.8	81.9	100.0	71.2
Morocco	119	23.4	57.1	53.5	78.5	99.6	65.3
Tunisia	94.2	61.8	94.8	94.6	97.7	100.0	74.4
Asia							
Indonesia	48.5	66.3	76.0	79.3	86.3	100.0	91.7
Sri Lanka	89.8	79.3	93.1	95.6	97.4	100.0	99.1
Thailand	88.7	74.8	90.7	91.8	96.0	99.5	85.7
Latin America/Caribbean							
Bolivia	96.6	83.0	95.6	96.5	98.8	98.7	86.0
Brazil	99.4	90.3	96.3	98.0	99.3	100.0	97.5
Colombia	98.9	88.4	98.1	98.9	99.6	100.0	98.4
Dominican Rep.	100.0	78.8	96.8	98.3	98.9	100.0	96.9
Ecuador	96.9	82.3	94.2	95.9	98.0	97.5	99.8
El Salvador	95.5	50.6	98.5[a]	98.1[a]	99.7[a]	100.0	95.8
Guatemala	96.3	77.8	96.2	96.6	99.3	100.0	97.4
Mexico	98.0	94.7	98.6	99.0	99.7	98.9	98.6
Peru	99.0	87.9	97.9	99.1	99.6	99.1	94.4
Trinidad & Tob.	99.8	75.4	99.0	99.3	99.7	99.5	98.9

[a] Truncated birth history covering the five plus years prior to the survey.
All data are unweighted.

B. Date of first union

In contrast to the date of birth of the respondent, the date of first union is much less well reported. Even in Latin America and the Caribbean where reporting of dates of birth are very high, the reporting of date of first union is considerably lower. This is for two main reasons:

Firstly, the start of a stable union is much less clearly defined than the start of a formal marriage or the date of a birth; and secondly, most DHS-I surveys collected the date of first union, but only asked for the age at first union if the year of first union was unknown.

Bolivia, Brazil, Dominican Republic, Ecuador, El Salvador, Peru, Indonesia, Togo and Thailand did collect both month and year of first union and age at first union in their surveys, but in all of these surveys there were considerable problems of consistency of date and age reporting of the first union.

Only in the surveys in Mali (24 percent) and Guatemala (9 percent) were there large numbers of respondents for whom no date or age at first union was recorded. In all other surveys 2 percent or less have no information. In cases where there is no information, there is a very wide range of possible dates of first union, with few constraints on the limits. In each of these cases it is assumed that the first union started before the first birth, however this assumption would clearly not be true in all cases. To constrain the lower bound of the date of first union, the age at first sexual intercourse is used, on the assumption that the first sexual intercourse takes place before or at the time of the start of the first union. Within these relatively wide bounds the date of first union was imputed.

C. Dates of birth of children

The dates of birth of children are generally well reported, and even in cases where the exact date was not specified, at least the year of birth of the child or its age was known. Thus the imputation process would be restricted to, at worst, a 12 month range, and usually a narrower range for the date of birth of the child. Only in Mali (10 percent) and Thailand (3 percent) were there many children recorded with no date or age information. The majority of cases with no date information were children who have died.

Recording of dates of birth of children was better, the nearer the birth was to the date of interview and children born in the five years prior to the survey generally have been recorded with both month and year of birth.

Differences between countries in the level of reporting of dates is partially due to the style of questionnaire used (Liberia, Ondo State, El Salvador) or the degree of training and instruction the interviewers received. In Uganda interviewers were trained to record dates of birth and ages for all children, even if it required manual imputation of dates during the interview.

D. Date of conception of current pregnancy

The date of conception of the current pregnancy is imputed from the duration of current pregnancy. Almost all respondents were able to report the duration of the pregnancy. The few cases were this was not true were usually because the interviewer failed to record the duration of pregnancy on the questionnaire rather than the respondent not knowing how long she had been pregnant.

E. Date of sterilization

The date of sterilization was only recorded for countries that used the DHS-I Model A questionnaire for high contraceptive prevalence countries. Only the date of the sterilization is recorded and not the age of the respondent at that date. In almost all of the cases where the full date of the sterilization is not known, at least the year of sterilization is recorded, restricting the imputation range to no more than 12 months.

V. Problem areas

During the processing of the DHS-I surveys, several problem areas appeared in the imputation process. Each of these problems is described below:

A. 5 year cut-off bias

Each of the DHS-I questionnaires has a cut-off for the inclusion of children for the questions in the health section of the questionnaire, set at January 1st of the year five years before the year of interview. The imputation process used this information in constraining the dates of birth of children. If there was information in the health section of the questionnaire, it was assumed that the child was born on or after this cut-off date, and if no information existed the date of birth of the child was assumed to be before this cut-off date. However, for children without a year of birth or an age, most interviewers tended to exclude them from the health section of the questionnaire. This was also re-enforced in some surveys by the data entry program making the assumption that children without a year of birth or age were born before the cut-off date.

These assumptions have lead to a bias in the imputation process, which has affected reported fertility and mortality rates (Arnold 1990, Sullivan et al. 1990). In most surveys the affect is slight as the number of cases with no year of birth or age of the child is generally small. A few surveys, though, show a significant bias caused by a number of factors, including the five year cut-off assumption in the imputation process. It should be noted, however, that the main reason for these biases is actually due to a tendency of the interviewers for dating the births of children earlier than the cut-off date for the health section, in situations where the exact date was not clear.

B. Year of birth calculation from age

In some surveys it became apparent that interviewers were calculating the year of birth of a child by subtracting the age of the child from the year of interview, or calculating the age of the child by subtracting the year of birth from the year of interview. The month of birth was usually left unknown when this took place. This calculation process led to a bias in the distribution of imputed dates of birth according to the month in which the birth took place, with significantly higher number of births in the months from January to the month of interview and lower in the months after the month of interview for each year.

In two countries, Mali and Ghana, adjustments were made to the imputation process to attempt to alleviate the problem. In Mali, for alternating births for which the month was unknown, but the year of birth and the age of the child was reported, either year of birth or the age of the child was ignored in the imputation process. This produced a less biased distribution of dates of birth.

In Ghana it was apparent during the fieldwork, that in the vast majority of these cases, it was the year of birth, which was being calculated from the reported age of the child (it is suspected that this was also true in Mali). For this reason, the year of birth of the child was ignored in the cases where the month of birth was unknown, but the year of birth and age were reported and the age plus the year of birth added up to the year of interview. Again this significantly reduced the bias, and it is believed that the resulting distribution better reflects the real situation.

This problem has only been seen in Sub-Saharan African countries, but that is due to the relatively lower levels of complete reporting of dates of events in these countries compared to the high levels found in other regions of the world.

C Ancillary data bias

Ancillary data, such as the durations of breastfeeding, amenorrhea and abstinence after the birth of a child and the duration of contraceptive use between two births, are used to constrain the bounds of the dates of births surrounding the intervals to which the ancillary data relates. These data are particularly prone to heaping. Cases where the ancillary data are rounded up produce greater constraints on the dates of births than may have been true in reality. Similarly data that are rounded down produce less of a constraint than in the real life situation. This bias is sometimes known as being "a half too smart" as this information is only ever used to narrow ranges for dates of events and not to enlarge the ranges. There is no obvious solution to this problem other than not using the ancillary data to constrain the dates of events and doing this would lead to the imputation of dates of events which may be inconsistent with the ancillary data reported.

D. Fine temporal variables

The process used in DHS to produce imputed dates of events from incomplete reporting has stood up well in most analyses and can be shown to have no significant bias on the results. However, for certain types of analysis, particularly those involving fine temporal variables where accuracy of reporting to the month is more important, there are some clear biases. The two most obvious cases are those relating to birth intervals and to pre-marital births.

The first problem arises when analyzing birth intervals. The distributions found in the data files appear to fairly represent the population when considering the aggregate level. However, when individual cases are investigated it is clear that some of the imputed dates of events are less than plausible given certain assumptions about spacing of births. A possible solution is to use midpoint imputation rather than random imputation in imputing the dates of events, but this may well introduce other biases. When carrying out birth interval analysis, and particularly the analysis of short birth intervals, a researcher should be aware that the short birth intervals may be a result of the imputation process and not necessarily the real situation.

The second problem concerns the proportion of premarital births in the surveys. When constraining the date of birth of the first child, the date of first union can be used as a constraint in order to avoid, if possible, premarital births being imputed, or the date of first union can be ignored allowing the imputation program to impute dates of birth prior to the date of first union. In either situation biases are found in the proportion of pre-marital births, being under estimated in the first case and over-estimated in the second case (Meekers 1991). In DHS-I the date of first union was generally ignored as a constraint on the date of first birth, except for the surveys in Islamic and Buddhist countries, that is, Egypt, Indonesia, Morocco, Sri Lanka, Sudan, Tunisia and Thailand.

E. First or last birth date unknown

In a small number of cases, the date of birth of the first child is completely unknown, that is**,** no month or year of birth and no age were specified for the child. The age at first sexual intercourse is usually used to constrain the date of birth, but this may either be unknown or may be at a very young age. In these cases the range of consistent dates for the date of birth of the child is extremely large and the imputation process may impute a date of first birth for the respondent at a very young age.

Similarly, if the last birth date and child's age information are not known and the preceding birth was a long time before interview, the range of possible dates for the last birth may be very large.

In this case the birth may sometimes be imputed with a much larger birth interval from the preceding birth than might be expected.

VI. Changes for DHS-II

The most significant change in the DHS-II date editing and imputation process has been to try to maintain the original data in the final data file when the data was found to be inconsistent This has been achieved by the inclusion of flag variables for several of the major variables that are often found to be inconsistent. The flag variables are set automatically as part of the imputation process as inconsistencies are found in the data. In the past these data values would have been changed manually to the inconsistent code (97) and the original data would have been lost.

Several other changes have been made in the date editing and imputation process for DHS-II in response to some of the problems outlined above.

A. 5 year cut-off

In DHS-I the existence of information in the health section for a particular child was used as a constraint on the date of birth of the child, restricting the birth date to before a certain date if no data existed in the health section, and after that date if data existed for the child. This constraint has been dropped in DHS-II due to the biases that it produced.

B. First union within event table

In DHS-II it was decided that the first union should be included in the event table as a constraint on the date of birth of the first child. This will tend to reduce the number of pre-marital births imputed in the data, but births that were clearly before the first union will remain so in the data. The effect of this change will probably be to under-estimate the number of pre-marital births, but it was felt that the previous procedures produced an over-estimate.

C. Allowance for miscalculation of year from age

Due to the problems found in some countries, particularly in Sub-Saharan Africa, where the interviewers appeared to be calculating the year of birth of a child from the age of the child, the DHS-II procedures have been adjusted to take this into account. A change has been incorporated in the DHS-II procedures to allow the year of birth to be ignored if the age of the child plus the year of birth add up to the year of interview <u>and</u> the month of birth of the child is unknown. This is the same modification made in the DHS-I survey in Ghana.

This allowance will be used in surveys in which the quality of date reporting is not so high. In countries where the level of complete date reporting is high, this adjustment is unlikely to be necessary.

D. Possible improvements

Several possible areas of improvement exist in the date editing and imputation procedures. The main areas being considered relate to the situations where improbable dates of events are imputed, and involves, firstly, imputing dates of birth with more realistic birth intervals between two births and, secondly, imputing dates of first birth and last birth that are more in line with the other births reported in the birth history.

A possible solution to the first case is to use midpoint imputation but, as mentioned above, this could lead to other biases, due to an averaging of birth interval durations, because fewer short or long birth intervals would be created. It should be remembered that it is dates that are imputed and not intervals, and that the imputation of one dale affects two intervals.

One idea, that requires further consideration, is to use a distribution of birth interval durations as a basis for the imputation, randomly assigning birth interval durations based on this distribution. The main problem with this method is that there are two intervals involved and that imputing a longer birth interval for one will produce a shorter birth interval for the other. This method, however, may be suitable to resolve the second situation in which there is, effectively, only one birth interval of importance, with the other being either the interval before the First birth or after the last birth.

VII. Conclusions

A. Summary

This paper has attempted to present the procedures used in the Demographic and Health Surveys program for data editing and imputation. It has described each stage of the imputation process in detail, with the intention of highlighting the complexity of the process and the need for procedures of this kind.

The procedures used in the DHS surveys are not without their problems. The major problems are discussed in the paper, with the hope that techniques for handling data of this kind can be improved. The problems are also presented to encourage the analyst to look carefully at the data before assuming that the imputation procedures always produce correct data. The main aim of date editing and imputation is to make data more readily usable and to allow the results of one analysis to be reproduced by another researcher, but not to mislead the analyst into thinking that the data are without problems.

B. Need for improved reporting

The process of executing a large-scale survey, has always been a complicated one, and quality of data has always been a major issue with all surveys. As techniques for survey data collection and for data processing have improved, so has the quality of data produced. The need for data editing and imputation techniques, though, serves to indicate that there is still a long way to go. There is still a need for better reporting of information and for better interviewing, with the hope that, one day, data editing and imputation techniques such as those presented will be redundant

VIII. References

Arnold. Fred. 1990. "Assessment of the Quality of Birth History Data in the Demographic and Health Surveys." Pp. 83-111 in *An Assessment of DHS-I Data Quality*. DHS Methodological Reports No. 1. Columbia, Maryland: Institute for Resource Development/Macro Systems Inc.

Chidambaram, V.C. and Z.A. Sathar. 1984. “Age and Date Reporting.” *WFS Comparative Studies* No. 5. Voorburg: International Statistical Institute.

Institute for Resource Development (IRD). 1987. *Demographic and Health Surveys Data Processing Manual.* Columbia, Maryland: IRD.

Meekers, Dominique. 1991. “The Effect of Imputation Procedures on First Birth Intervals: Evidence from Five African Fertility Surveys.” *Demography* 28(2): 249-260.

Otto, James. 1980. “Date Imputation and Recode Program: User's Manual.” *WFS Technical Paper* No. 1430 (Unpublished mimeographed document).

Otto, James and Judith Rattenbury. 1987. “WFS Data Processing Strategy.” Pp. 477-514 in *The World Fertility Survey: An Assessment* edited by John Cleland and Chris Scott. London: Oxford University Press.

Sullivan, Jeremiah M., George T. Bicego and Shea O. Rutstein. 1990. “Assessment of the Quality of Data Used for the Direct Estimation of Infant and Child Mortality in the Demographic and Health Surveys.” Pp. 115-137 in *An Assessment of DHS-I Data Quality*. DHS Methodological Reports No. 1. Columbia, Maryland: Institute for Resource Development/Macro Systems Inc.

Trussell, James. 1987. “Date Imputation.” Pp 677-712 in *The World Fertility Survey: An Assessment* edited by John Cleland and Chris Scott. London: Oxford University Press.

World Fertility Survey. 1980. “Data Processing Guidelines.” *WFS Basic Documentation* No. 11 volumes 1 and 2.